# Somatic Diversification of Rearranged Antibody Gene Segments by Intra- and Interchromosomal Templated Mutagenesis

Gordon A. Dale,* Daniel J. Wilkins,* Jordan Rowley,† Christopher D. Scharer,‡ Christopher M. Tipton,§ Jennifer Hom,§ Jeremy M. Boss,‡ Victor Corces,† Ignacio Sanz,§ and Joshy Jacob*

**The ability of the humoral immune system to generate Abs capable of specifically binding a myriad of Ags is critically dependent on the somatic hypermutation program. This program induces both templated mutations (i.e., gene conversion) and untemplated mutations. In humans, somatic hypermutation is widely believed to result in untemplated point mutations. In this study, we demonstrate detection of large-scale templated events that occur in human memory B cells and circulating plasmablasts. We find that such mutations are templated intrachromosomally from IGHV genes and interchromosomally from IGHV pseudogenes as well as other homologous regions unrelated to IGHV genes. These same donor regions are used in multiple individuals, and they predominantly originate from chromosomes 14, 15, and 16. In addition, we find that exogenous sequences placed at the IgH locus, such as LAIR1, undergo templated mutagenesis and that homology appears to be the major determinant for donor choice. Furthermore, we find that donor tracts originate from areas in proximity with open chromatin, which are transcriptionally active, and are found in spatial proximity with the IgH locus during the germinal center reaction. These donor sequences are inserted into the Ig gene segment in association with overlapping activation-induced cytidine deaminase hotspots. Taken together, these studies suggest that diversity generated during the germinal center response is driven by untemplated point mutations as well as templated mutagenesis using local and distant regions of the genome.** ***The Journal of Immunology*****, 2022, 208: 2141–2153.**

To generate an effective humoral immune response, diversity in Ag-specific Abs is paramount. Such diversity is used as a raw substrate in the germinal center (GC) reaction in which high-affinity Abs are selected at the expense of their weaker-binding kin (1). For this Darwinian microcosm to proceed effectively, GC B cells divide and mutate rapidly, allowing for a new round of selection to occur. Repeated cycles of this process lead to the characteristic high-affinity Abs of the humoral immune response.

To generate diversity between cycles of the GC reaction, GC B cells engage in a mutagenic program known as somatic hypermutation (SHM) that is chiefly mediated by activation-induced cytidine deaminase (AID) (2–5). This enzyme controls two distinct pathways of mutagenesis: canonical SHM and gene conversion (2, 3). In canonical SHM, AID targets the Ab loci and deaminates cytosines that are then processed by multiple repair enzymes leading to de novo somatic mutations (4, 5). In gene conversion, activity of the AID enzyme and subsequent processing can lead to a form of mutagenic repair that copies homologous sequences, resulting in mutations templated from such sequences (3, 6).

Although it is known that gene conversion can occur in murine and human B cells, this process is often regarded as infrequent, and exact frequencies of gene conversion remain unknown (7–9). Prior approaches to quantitate the contribution of gene conversion to SHM have suffered from reliance on imprecise measures of gene conversion tracts, limitations on what sequences could serve as gene conversion donors, as well as false positives (7, 10–13). In this study, we present data to show the genome-wide contribution of templated mutagenesis (i.e., gene conversion) using sequences from LAIR1-insert Abs and human donors. We do this via a novel computational script known as template recognition via Monte Carlo experiments (TRACE) that relies on the BLASTn (14) script to identify likely gene conversion donors. We then supplement our findings with published human GC B cell data, showing that TRACE-predicted gene conversion donors are predictive of open chromatin peaks, RNA transcription, and chromosomal orientation of GC B cells.

We report in the present study that TRACE identifies templated mutations in both somatically mutated, rearranged IGHV gene segments as well as in non-Ig sequences at the IgH locus. These mutations are found to account for 0.7% of the total mutation load.

*Emory Vaccine Center, Yerkes National Primate Center, Emory University, Atlanta, GA; †Department of Biology, Emory University, Atlanta, GA; ‡Emory University School of Medicine, Emory University, Atlanta, GA; and §Lowance Center for Human Immunology, Department of Medicine, Emory University, Atlanta, GA

Received for publication May 10, 2021. Accepted for publication February 15, 2022.

ORCIDs: 0000-0002-4995-6017 (G.A.D.); 0000-0003-0398-8715 (D.J.W.); 0000-0002-5135-9596 (J.R.); 0000-0001-7716-8504 (C.D.S.); 0000-0002-5049-2571 (C.M.T.); 0000-0002-2432-1840 (J.M.B.); 0000-0001-5140-4337 (V.C.); 0000-0003-0247-7016 (J.J.).

This work was supported by National Institute of Allergy and Infectious Diseases Grants U19 AI117891 and F30 AI124568.

Conceptualization, G.A.D. and J.J.; methodology, G.A.D., D.J.W., J.R., V.C., C.D.S., J.M.B., I.S., and J.J.; investigation, G.A.D., D.J.W., J.R., C.M.T., and J.H.; visualization, G.A.D.; funding acquisition, G.A.D. and J.J.; project administration, G.A.D. and J.J.; supervision, J.M.B., V.C., I.S., and J.J.; writing – original draft, G.A.D.; writing – review & editing, G.A.D. and J.J.

The data presented in this article have been submitted to the Sequence Read Archive under accession numbers SRR17118783–SRR17118786.

Address correspondence and reprint requests to Dr. Joshy Jacob, Emory University, Yerkes National Primate Research Center, 954 Gatewood Road NE, Atlanta, GA 30329. E-mail address: jjacob3@emory.edu

The online version of this article contains supplemental material.

Abbreviations used in this article: AID, activation-induced cytidine deaminase; FPR, false positive rate; GC, germinal center; RNA-seq, RNA sequencing; RPKM, reads per kilobase transcript per million mapped reads; SHM, somatic hypermutation; SLE, systemic lupus erythematosus; TRACE, template recognition via Monte Carlo experiments; 10x, 10x Genomics; *Zs*, Stouffer's *Z* score.

www.jimmunol.org/cgi/doi/10.4049/jimmunol.2100434

Templates identified by TRACE cluster intrachromosomally and interchromosomally between individuals. Analysis of these regions demonstrate that TRACE donors are in areas in proximity to open chromatin, which are transcriptionally active, and are in spatial association with the IgH locus during the GC reaction. Furthermore, TRACE-identified tracts in somatically mutated genes are found to be in association with overlapping AID hotspots. We also validate TRACE's outputs through detailed false-positive and technical artifact analysis.

## Materials and Methods

### *TRACE script*

TRACE is a custom script written in MATLAB (v.2018a) and uses nested, iterative BLASTn to identify donor templates for somatically mutated sequences at the genome scale. FASTA files containing a germline reference and somatically mutated sequences are parsed for user-defined mutation clusters (in this study, eight or more mutations over 38 bp). A preprocessing step is included to remove insertion/deletion events from the analyzed sequence sets. Subsequences containing the mutation cluster are split into 38-bp windows that each contain eight or more mutations. Each of these windows is passed into BLASTn (word size, 11; maximum high-scoring segment pairs, one; maximum target sequences, one) against either the human genome (GRCh38) or the mouse genome (GRCm38). The window with the greatest bit score is stored and passed into two sequential Monte Carlo analyses, with entry into the second contingent on the results in the first. In the first Monte Carlo analysis, the effect of the mutation's identity in the window is assayed by randomizing the identity of the mutated bases, generating a window with different combinations of mutations. One thousand simulated windows are passed into BLASTn and their respective bit scores are used to build a population to which the original stored window is compared. If the $Z$ score of the original window is $\geq 1.645$, then the window is passed into the second Monte Carlo analysis, wherein the effect of the location of mutations is assayed. In this study, the numbers of mutations in the original window are randomly shuffled over the length of the window. As before, 1000 simulated windows are generated and passed into BLASTn to generate a second population of bit scores. If the $Z$ score of the original window is $\geq 1.645$ as compared with this second population, the original BLAST hit is stored as a TRACE hit. This process is iterated through all sequences until all mutation clusters have been analyzed.

Alongside the analysis of the input dataset, a series of 10 "background" datasets are generated in which the original FASTA file is analyzed for mutation clusters as above. The number of mutation clusters and the corresponding number of mutations per cluster per sequence is then randomized such that the locations of the mutation clusters are randomly placed along the length of the sequence. Each of these background sets undergoes the same core analysis above involving cluster identification, subsequent Monte Carlo analyses, and recording of TRACE hits that pass both Monte Carlo analyses.

Upon completion of analysis of the background data and the original dataset, TRACE hits are passed through BLAST and the corresponding template is locally aligned. The number of mutations accounted for by the top BLAST hit is calculated and stored for each of the TRACE hits in the modeled and original datasets. Other data are also gathered at this time, including the percent identity between the TRACE hit and its corresponding template, the strand to which the hit localizes, the gene name (if any), and whether the identified template is an exon or intron.

Finally, TRACE hits from the original data are compared with that of the background by placing the data in bins composed of the combination of length, the number of mutations explained by the TRACE-identified template, and the percent identity of the mutation cluster window to the identified template. The frequency of hits in each bin is counted for the original and the background datasets. Any bins unique to the original data are defined as true hits. For bins in both datasets, original data TRACE hits are only retained when the frequency of the original bin is greater than the frequency of the modeled bin plus 2 SD of all the frequencies of bins in the modeled dataset. Reported data from TRACE are cleaned such that overlapping TRACE hits that map back to the same template are removed.

### *Subjects*

Three healthy subjects vaccinated with trivalent influenza vaccine (701, 702, 752) and one systemic lupus erythematosus (SLE) patient experiencing an acute flare (730) were enrolled in this study at Emory University between 2013 and 2015. Cell populations examined include class-switched memory B cells (701, 702) or circulating Ab-secreting cells (730, 752). Healthy subjects received the influenza vaccine as part of routine medical care. SLE patient recruitment outside the annual influenza season and patient history were used to determine absence of recent immunization or likely natural exposure to influenza. PBMCs were isolated on days 6–9 for vaccination subjects. All studies were approved by the Institutional Review Boards at Emory University School of Medicine. The SLE patient fulfilled four or more criteria of the modified American College of Rheumatology classification and was routinely evaluated by expert rheumatologists at the Emory Lupus Clinic. The SLE patient was classified as having a moderate–severe flare according to the SELENA-SLEDAI flare index and were on minimal immunosuppression at the time of flare (only hydroxychloroquine and/or <10 mg/d prednisone or equivalent glucocorticoid).

Sequences acquired from 10x Genomics (10x) were downloaded from the datasets page (https://www.10xgenomics.com/resources/datasets). The seven datasets used in this study are as follows: 1) human B cells from a healthy donor, 1000 cells: Multi (v2); single-cell immune profiling dataset by Cell Ranger 5.0; 2) non–small cell lung cancer tumor, single-cell immune profiling dataset by Cell Ranger 5.0; 3) PBMCs of a healthy donor (Next GEM v1.1), single-cell immune profiling dataset by Cell Ranger 3.1.0; 4) human PBMCs from a healthy donor, 10,000 cells (Multi v2), single-cell immune profiling dataset by Cell Ranger 4.0.0; 5) PBMCs of a healthy donor (Next GEM v1.1, 150 × 150), single-cell immune profiling dataset by Cell Ranger 3.1.0; 6) PBMCs of a healthy donor: Ig enrichment from amplified cDNA, single-cell immune profiling dataset by Cell Ranger 3.0.0; and 7) human B cells from a healthy donor before and after flu vaccination (Multi v2), single-cell immune profiling dataset by Cell Ranger 5.0.0.

### *Multicolor flow cytometry and sorting*

Mononuclear cells were isolated from peripheral blood using Ficoll density gradient centrifugation and stained with the following anti-human Ab staining reagents: IgD-FITC, CD3-Pacific Orange, CD14-Pacific Orange, CD24-PE-A610 (Invitrogen, Camarillo, CA); CD19-allophycocyanin-Cy7, CD38-Pacific Blue, CD23-PE-Cy7, CD21-PE-Cy5, CD27-PE (BD Pharmingen, San Diego, CA); and CD138-allophycocyanin (Miltenyi Biotec, Auburn, CA). Approximately 30,000 cells were collected for either switched memory B or plasmablast populations using a BD FACSAria II (BD Biosciences, San Jose, CA) and sorted directly into RLT lysis buffer (Qiagen, Valencia, CA).

### *Next-generation sequencing of the IgH repertoire*

Total cellular RNA was isolated from each sample using the RNeasy Micro kit by following the manufacturer's protocol (Qiagen, Valencia, CA). Approximately 2 ng of RNA was subjected to reverse transcription using the iScript cDNA synthesis kit (Bio-Rad, Hercules, CA). Aliquots of the resulting single-stranded cDNA products were mixed with 50 nM VH1–VH7 FR1-specific primers and 250 nM Cα-, Cμ-, and Cγ-specific primers preceded by the respective Illumina Nextera sequencing tag (sequences listed below) in a 25-μl PCR reaction (using 4 μl of template cDNA) using Invitrogen's high-fidelity Platinum PCR SuperMix (Invitrogen, Camarillo, CA). Amplification was performed with a Bio-Rad C1000 thermal cycler (Bio-Rad, Hercules, CA) with the following conditions: PCR1: 95°C for 5 min; 35 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s; 72°C for 5 min.

A second PCR was used to add Nextera indices with the following conditions: PCR2: 72°C for 3 min, 98°C for 30 s; five cycles of 98°C for 10 s, 63°C for 30 s, and 72°C for 3 min.

Ampure XP beads (Beckman Coulter Genomics, Danvers, MA) were used to purify the products, and they were subsequently pooled and denatured. Single-strand products were sequenced on a MiSeq (Illumina, San Diego, CA) using the 300 bp ×2 v3 kit. Primers for PCR1 were as follows: forward: VH1a, 5′-CAGGTKCAGCTGGTGCAG-3′; VH1b, 5′-SAGGTCCAGCTGGTACAG-3′; VH1c, 5′-CARATGCAGCTGGTGCAG-3′; VH2, 5′-CAGGTCACCTTGARGGAG-3′; VH3, 5′-GGTCCCTGAGACTCTCCTGT-3′; VH4, 5′-ACCCTGTCCCTCACCTGC-3′; VH5, 5′-GCAGCTGGTGCAGTCTGGAG-3′; VH6, 5′-CAGGACTGGTGAAGCCCTCG-3′; VH7, 5′-CAGGTGCAGCTGGTGCAA-3′; reverse: Cμ, 5′-CAGGAGACGAGGGGGAAAAGG-3′; Cγ, 5′-CCGATGGGCCCTTGGTGGA-3′; Cα, 5′-GAAGACCTTGGGGCTGGTCG-3′; F tag, 5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3′; R tag, 5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3′.

### *Bioinformatics analysis of next-generation sequencing data*

An in-house–developed informatics pipeline was used for initial quality filtering and clonal clustering analysis of sequencing data. After paired-end reads were joined, sequences were filtered based on a length and quality threshold. Sequences <200 bp and sequences with poor overlaps (>8% difference in linked region) and/or a high number of base pairs below a threshold score (sequences containing >15 bp with <Q30 score, 10 bp with Q20 score, or

any base pair with <Q10 score) were excluded from further analysis. Isotypes were then determined by analysis of the C region segment of each sequence, and then sequences were aligned using the data provided by IMGT/HIGHV-QUEST (http://www.imgt.org/HIGHV-QUEST/) (15). See Tipton et al. 2015 (16) for further reasoning and analysis.

### *Overlapping and nonoverlapping AID hotspot analysis*

Analysis of overlapping AID hotspots was performed by identifying locations that contain the WGCW motif, where the mutated base is underlined. As these sites are palindromic, locations for these hotspots were counted twice, once for each strand. To determine whether TRACE recipient sites were located proximally to overlapping hotspots, the average shortest distance between a given TRACE recipient site and any given overlapping AID hotspot was calculated. To determine significance, each dataset had the location of TRACE recipient sites randomized across the length of the sequence, and the average shortest distance was calculated over 1000 iterations. *Z* scores for each individual dataset were determined in comparison with each sequence set's randomized set. *Z* scores were combined into a single statistic using Stouffer's *Z* method. This method was repeated for analysis of WRC motifs, which were defined as WRC motifs that did not include any WGCW sites. Modeled sets were not included in this analysis, as clusters of mutations in these sets were randomly distributed across the span of the IGHV sequence.

### *Clustering analyses*

Clustering was performed either on the chromosomal level or on the base pair level and done using a permutational or Monte Carlo approach, respectively. For permutational approaches, equal numbers of TRACE hits were randomly assigned to a chromosome with weights to adjust for differences in chromosome size. The frequency of TRACE hits per chromosome bin were tallied and recorded over 1000 iterations, and the frequency of TRACE hits at their original positions were compared with the permuted pool for each chromosome. *Z* scores were determined by comparing the original TRACE hit frequency to that of each chromosome in the permutated pool. *Z* scores were combined into a single statistic using Stouffer's *Z* method.

For Monte Carlo approaches, equal numbers of TRACE hits were randomly assigned to specific locations in the genome, with weights given to account for differences in chromosome size, as above. Counts were determined for TRACE hits that occurred within 1 kb of each other over 1000 iterations of randomly assigned TRACE hits. Counts were also determined for original TRACE hits, and the original count was compared with the population of counts from the randomly assigned pool.

In each approach, modeled sets were used to control for the innate background generated from the TRACE methodology. These modeled sets were matched to each sequence set analyzed and control for the number of mutations and clusters found in each set. Each modeled set was run through TRACE with identical settings to the original data.

### *GC transcript analysis*

GC B and naive B cell RNA sequencing (RNA-seq) data were obtained from GSE84022 (17), and reads per kilobase transcript per million mapped reads (RPKM) values were averaged between replicates. GC B cell–upregulated transcripts were defined by having a RPKM value greater than that of naive B cells plus 2 SD. Downregulated transcripts were similarly defined, except that the RPKM value for transcripts in GC B cells was less than that of naive B cells minus 2 SD. Permutational analysis was done by randomly selecting an equal number of genes and assaying how many genes in this set were present in either fraction for 1000 iterations. Original values were compared with those generated during the permutation process to yield *Z* scores. RPKM values associated with each gene were used in the analysis to yield *Z* scores for RPKM values.

### *Hi-C analysis*

GC B and naive B cell Hi-C data were obtained from GSE84022 (17), and interchromosomal interactions were kept when one anchor overlapped the IgH locus and was supported by at least five reads, representing interactions >98% of IgH interacting pairs. Cumulative plots of interchromosomal interactions were obtained by taking each examined locus and the closest distance to an anchor that was found to interact with the IgH locus. These distances in kilobases were $\log_{10}$ transformed and plotted with an empirical cumulative distribution function. Kolmogorov–Smirnov tests were used to test the significance of the observed differences in the distributions between samples.

### *Statistical analysis*

Tests for significance used in this study include a paired *t* test, Tukey post hoc test, one-sample *Z* test, Stouffer's *Z* method, and a Kolmogorov–Smirnov test. Paired *t* tests were used when comparing percent mutations and *Z* scores between original and matched modeled sets. A Tukey post hoc test was used in analysis of TRACE output types. For *t* tests and the Tukey post hoc test a two-tailed α of 0.05 was used. One-sample *Z* tests were used in either permutational or Monte Carlo analyses. For all tests, significance was set at a *Z* score of 1.645, which corresponds to a one-tailed α of 0.05. In cases where multiple *Z* scores were combined for an aggregate statistic, only samples belonging to the same group were combined, and Stouffer's *Z* method was used. As in the one sample *Z* test, a cutoff *Z* score of ±1.645 was used to determine significance, with direction being chosen depending on the analysis. Kolmogorov–Smirnov tests were performed with a two-tailed α of 0.05.

### *Data and materials availability*

All data used in this study are available from the Sequence Read Archive under accession numbers SRR17118783–SRR17118786 (https://www.ncbi.nlm.nih.gov/sra). The TRACE source code is available at https://github.com/GDale1/Templated-Mutation-Detection-with-TRACE as well as at https://zenodo.org/record/5759792#.Yaz-5dDMKUk.

## Results

### *Somatic mutations in LAIR1 inserts display similarity to other genomic regions*

We have previously demonstrated that small clusters of mutations (two or more mutations over 8 bp) in somatically mutated sequences at the IgH locus are consistent with templated events (12). Our analyses have also demonstrated that the somatic mutation clusters present in the broadly neutralizing anti-malarial LAIR1-containing Abs reported by Pieper et al. (18) and Tan et al. (19) (Fig. 1A) are consistent with templated events. These LAIR1-containing Abs are atypical Abs that result from a templated insertion event of a segment of the LAIR1 gene on chromosome 19 into the CDR3 of an Ab rearrangement on chromosome 14 and are subsequently mutated to gain broad anti-malarial binding capacity. Interestingly, we observed that although many clusters of mutations in somatically mutated LAIR1 appeared to have templates corresponding to IGHV genes, there were multiple instances of heavily mutated regions that did not have a corresponding IGHV template. We performed BLASTn searches on these subsequences and found that these sequences had matches to distant genomic regions (Fig. 1B). Local alignment of resultant matches with the somatically mutated LAIR1 segment and the LAIR1 germline subsequence revealed that these matches account for several mutations found in the mutated subsequence and retain general homology to the LAIR1 segment (Fig. 1B). This raised the possibility that these clustered somatic mutations may derive from distant genomic regions but did not answer whether such matches were significant. Indeed, it remained a likely possibility that somatic mutations themselves could produce spurious alignments given the size and scale of the human genome.

### *TRACE analysis of LAIR1-containing Ab sequences*

To address this problem, we generated a custom MATLAB script called TRACE (Fig. 1C). This script was designed to determine whether clusters of mutations of a given density are statistically likely to have occurred through templated mutagenesis. In brief, TRACE operates through a nested system of Monte Carlo simulations that identifies clustered mutations (defined as at least eight mutations over 38 bp), and subsequently iterates and analyzes outputs of the BLASTn algorithm (Supplemental Fig. 1). The script conducts simulations to determine 1) whether the identity of the mutations within the cluster is important to produce a distant alignment, 2) whether the local position of the mutations relative to one another is important to produce the distant alignment, and 3) whether the resultant alignments are statistically different from if the germline sequence were randomly mutated. Only results that passed all three tests were considered for further analysis.

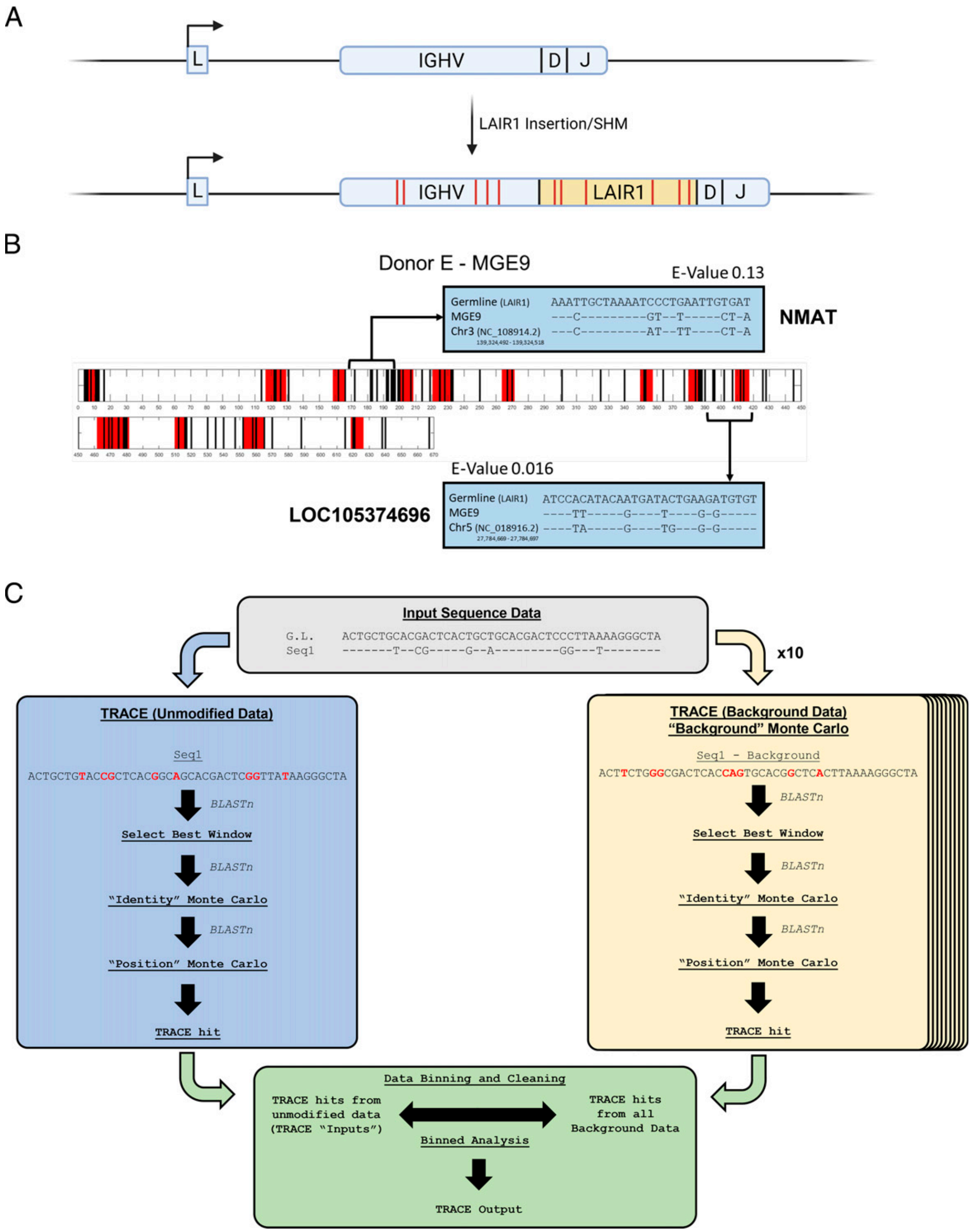

A
L
IGHV
D
J
LAIR1 Insertion/SHM
L
IGHV
LAIR1
D
J
B
Donor E - MGE9
E-Value 0.13
Germline (LAIR1) AAATTGCTAAAATCCCTGAATTGTGAT
MGE9 ---C----------GT--T-----CT-A
Chr3 (NC_108914.2) ---C----------AT--TT----CT-A
139,324,492 - 139,324,518
NMAT
E-Value 0.016
LOC105374696
Germline (LAIR1) ATCCACATACAATGATACTGAAGATGTGT
MGE9 ----TT-----G----T----G-G-----
Chr5 (NC_018916.2) ----TA-----G----TG---G-G-----
27,784,669 - 27,784,697
C
Input Sequence Data
G.L. ACTGCTGCACGACTCACTGCTGCACGACTCCCTTAAAAGGGCTA
Seq1 -------T--CG-----G--A---------GG---T--------
x10
TRACE (Unmodified Data)
Seq1
ACTGCTGTACCGCTCACGGCAGCACGACTCGGTTATAAGGGCTA
BLASTn
Select Best Window
BLASTn
"Identity" Monte Carlo
BLASTn
"Position" Monte Carlo
TRACE hit
TRACE (Background Data)
"Background" Monte Carlo
Seq1 - Background
ACTTCTGGGCGACTCACCAGTGCACGGCTCACTTAAAAGGGCTA
BLASTn
Select Best Window
BLASTn
"Identity" Monte Carlo
BLASTn
"Position" Monte Carlo
TRACE hit
Data Binning and Cleaning
TRACE hits from unmodified data (TRACE "Inputs")
TRACE hits from all Background Data
Binned Analysis
TRACE Output

**FIGURE 1.** Somatically mutated LAIR1 inserts have regions of clustered mutations that match distant genomic regions. (**A**) Schematic of the generation of LAIR1-containing Abs described by Tan et al. (19) and Pieper et al. (18). Germline V(D)J rearrangements acquire LAIR1 insertions in CDR3 that are somatically diversified and confer Ag binding. (**B**) Somatically mutated sequences obtained from human donor E were analyzed for small clusters of templated mutagenesis events as in Dale et al. (12) and are depicted as a highlighter plot. In the panel, a single sequence is shown; black bars indicate mutations at a given position, whereas red regions indicate regions that match short subsequences in the IGHV germline repertoire. Selected subsequences containing regions of clustered mutations (indicated by brackets) were run through BLASTn. Alignments of germline LAIR1, somatically mutated LAIR1, and the top BLASTn hit identified for that subsequence are shown. Gene names and E-values for searches conducted are shown for corresponding BLASTn hits. (**C**) An overview of the TRACE pipeline is shown. For each sequence, local clusters of mutations are identified by comparison with a germline sequence. Sequences with a qualifying cluster (eight mutations in 38 bp) are screened for the best window over the 38 bp that incorporates the cluster. Each of these windows is passed into BLASTn and the window with the highest bit score (i.e., the best alignment) is selected. Subsequently, two different Monte Carlo simulations are performed on this window and the bit score of the window is compared with the populations of bit scores from the Monte Carlo simulations. If the bit score of the window is >95% of each Monte Carlo population, the BLASTn result of the window is recorded. Simultaneously, 10 "background" datasets are generated that mimic the somatically mutated sequence, in both number of mutations and mutation clusters. Identical window selection and Monte Carlo simulations are performed with these background data. Significant BLASTn results are recorded. Data from the original sequence and the background sets are compared and BLASTn hits from the original sequences that exceed background (background data) are kept to produce a TRACE output. A detailed explanation of each step is depicted in Supplemental Fig. 1.

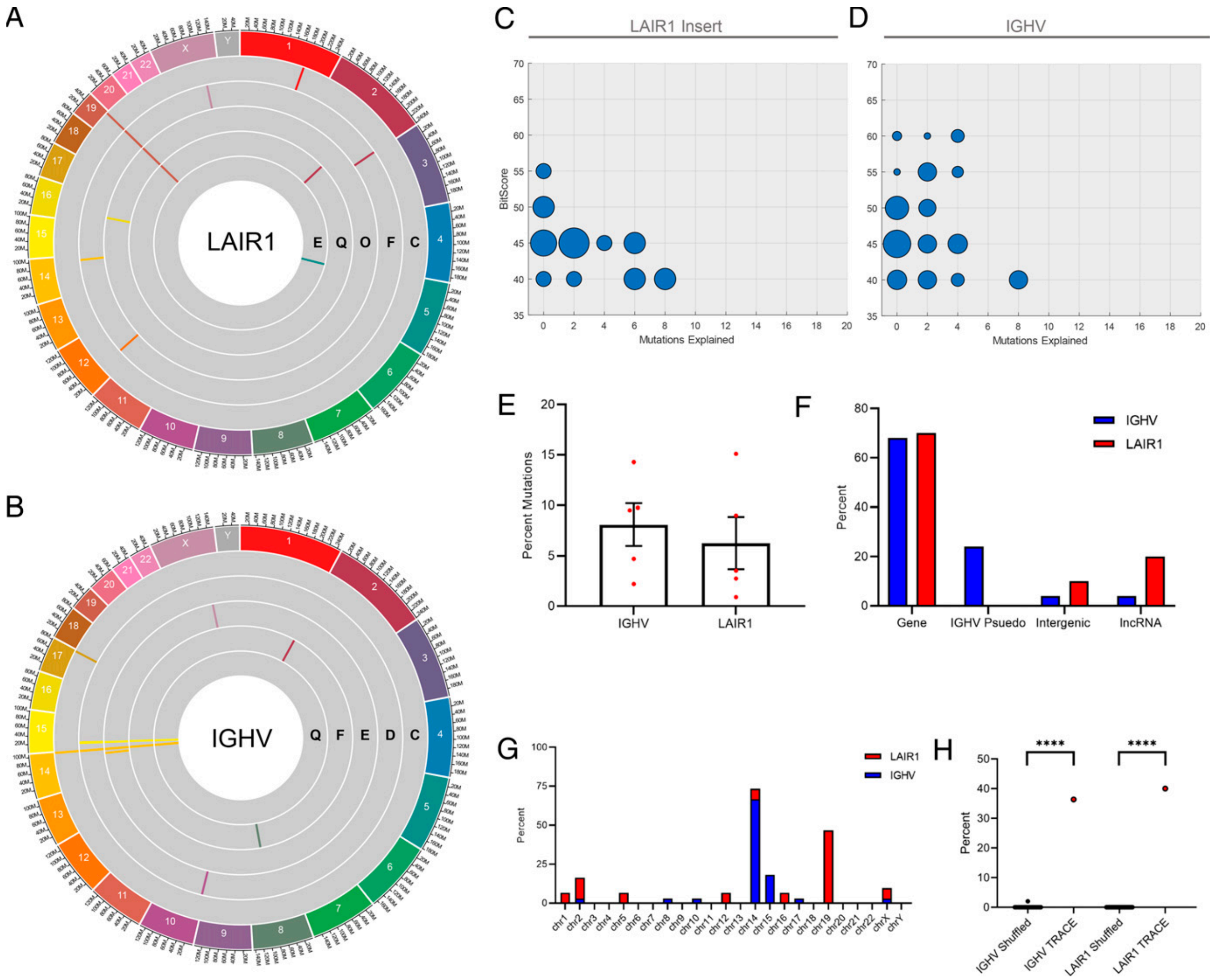


**FIGURE 2.** TRACE-identified regions of the genome contribute to somatic mutagenesis of IGHV/LAIR1 sequences. (**A** and **B**) Circos plots for (A) LAIR1 and (B) corresponding IGHV of multiple human donors (gray concentric circles). Human donors are annotated for each concentric circle and donor names correspond to those in Tan et al. (19) and Pieper et al. (18) (**C** and **D**) Summary data of TRACE outputs from LAIR1 (C) and IGHVs (D). Shown are density plots depicting the relative frequency of TRACE outputs as a function of alignment quality (bit score) and the number of explained mutations in the somatically mutated motif (mutations explained). Size of data points in the plot depict relative frequency compared with other data points in the plot. Larger points indicate higher frequency. (**E**) Shown are the percent of mutations that derive from a TRACE-identified donor sequence in either LAIR1 or IGHVs. (**F**) Bar plot depicting the types of TRACE-identified donor tracts in the LAIR1 or IGHV gene segments. (**G**) Stacked bar graph of the percent of TRACE hits per chromosome. (**H**) Percent of an equal number of randomly scattered TRACE hits (shuffled, $n = 1000$) and of the data generated from the LAIR1 rearrangements that cluster within 1 kb of one another. ****$p < 0.0001$.

Application of the TRACE pipeline to the LAIR1 and IGHV sequences present in Pieper et al. (18) and Tan et al. (19) revealed significant donor templates scattered around the genome (Fig. 2A, 2B). Notably, we observed that LAIR1 TRACE outputs localized to chromosome 19 in four donors and, similarly, IGHV TRACE outputs localized to chromosome 14 in all five donors, and to chromosome 15 in four donors. To describe the quality of these TRACE outputs, we considered the following two characteristics to be most relevant: the bit score assigned to the TRACE output, and the number of mutations within a given region explained by TRACE. The bit score is a product of the original BLASTn search and is reflective of the quality of the alignment. Thus, an ideal TRACE output is one with a high bit score and number of mutations explained. These outputs are most suggestive of gene conversion.

TRACE analysis of the LAIR1-containing Abs revealed a total of 48 outputs for both the somatically mutated LAIR1 insert as well as the corresponding IGHV segments. These outputs varied in quality, with high-quality TRACE outputs being relatively rare (Fig. 2C, 2D). We observed that TRACE outputs were generally higher in quality in the IGHV segment as compared with LAIR1 outputs. These TRACE outputs together accounted for 2–15% of mutations present in the somatically mutated LAIR1 and 3–15% of mutations for the somatically mutated IGHV segments (Fig. 2E). Analysis of the TRACE outputs revealed that gene-encoding regions were the primary contributor (Fig. 2F). Notably, IGHV pseudogenes were detected by TRACE as gene conversion donors but were restricted only to somatically mutated IGHV segments.

TRACE outputs were observed to cluster by chromosome for the LAIR1 and IGHV segments. Interestingly, donor templates clustered by chromosome with 22 out of 33 (66.7%) donors for the IGHV mutations originating intrachromosomally from chromosome 14 ($Z = 19.9$, $p < 0.001$) and 7 out of 15 (46.7%) donors for the LAIR1 mutations originating interchromosomally from chromosome 19 ($Z = 13.2$, $p < 0.001$) (Fig. 2G). Further analysis revealed that donor tracts not only clustered at the chromosomal level, but also at the nucleotide level, with ~40% of all TRACE outputs for LAIR1 derived from independent samples exhibiting significant clustering within 1 kb of one another ($Z = 189.7$, $p < 0.001$) (Fig. 2H). These clustered regions of donors were strikingly different depending on whether the recipient sequence was the LAIR1 insert or the IGHV gene segment, suggesting that sequence homology was the main determinant of donor sequences. Our observation that these donor templates cluster within a narrow genomic region and account for mutations across multiple samples suggested that there are preferred sites that serve as donor sequences.

### *False-positive analysis of TRACE pipeline*

Although these results were suggestive of templated mutagenesis, it was possible for these results to be the product of false positives, as had been shown to be of concern during large-scale analyses such as these by Fukuyama et al. (13). This was especially of concern given that TRACE outputs clustered on chromosomes where each segment is found (e.g., IGHV donors were found on the chromosome containing germline IGHVs). To address the contributions of false positives we investigated the somatically mutated Ab repertoire of bulk-sequenced class-switched memory B cells ($CD19^{+}IgD^{-}CD27^{+}$) or circulating Ab-secreting cells ($CD19^{+}IgD^{-}CD27^{hi}CD38^{hi}$) from four human donors. Additionally, to assess and control for PCR crossover, we also investigated seven public single-cell datasets provided by 10x and filtered sequences for high-confidence reads that were unlikely to be due to PCR or another artifact. These were chosen over the LAIR1 Abs due to the small sample size of the LAIR1 sequences compared with that of the memory pool of each donor and the 10x dataset (average number of analyzed sequences per donor $n = 9.67$, 66,137, and 15,268, respectively).

To accurately model for false positives, we leveraged the background set that is generated during each TRACE run. The background set mimics the input sequence set in terms of sequence number, mutation load, and number of local mutation clusters analyzed by TRACE but differs in where each mutation cluster is located and the identity of mutations within a given cluster. Thus, the background set retains key characteristics of the original dataset but should not produce significant TRACE-identified mutation donors. Therefore, any TRACE outputs generated by a model set that is passed through the TRACE analysis are, by definition, false positives.

From the four human donors, we selected three IGHV datasets (IGHV1−18, IGHV3−15, and IGHV5−51) from three donors ($n = 9$, total sequences analyzed $n = 14{,}378$). For each analysis, the original dataset was run in parallel with a background dataset generated during TRACE analysis of the original (Supplemental Fig. 2), which we refer to as the modeled set. In each of these modeled datasets, we obtained TRACE outputs indicating the contributions of false positives to our script. Given that the TRACE analysis pipeline relies on cleaning spurious outputs during the binning phase (Supplemental Fig. 1), we hypothesized that false positives would be a function of the total number of binned inputs. Analysis reveals that the false positive rate (FPR), defined as the percent of resultant TRACE outputs following binning, follows a power function (Fig. 3A). This indicates that false positives are high when there are low numbers of inputs during binning and vice versa. Comparison of the number of inputs for binning between the modeled set and their corresponding human donor data yielded no significant difference but trended toward a significant increase ($p = 0.0543$) (Fig. 3B), suggesting a lower FPR across the human donor data.

Subsequent analysis of the resulting TRACE outputs from the human donors and the corresponding modeled sets reveals that the quality of the TRACE outputs is increased in the human donor data (Fig. 3C, 3D). To ensure that the signal produced by TRACE was not the product solely of a PCR artifact secondary to bulk amplification and sequencing, we analyzed the TRACE outputs from the entirety of the 10x dataset and its corresponding modeled sets (Fig. 3E, 3F). As in the data from bulk sequencing, we detected an increase in TRACE output quality as compared with the modeled set, suggestive of a contribution of gene conversion independent of bulk versus single-cell processed samples. In both original datasets, we observed a linear trend of increasing bit score and number of mutations explained that is not observed in either modeled set. Interestingly, we also observed a subset of data in which large numbers of mutations are explained.

Given our analysis of false positives, we sought to characterize the expected rate of false positives from TRACE analysis of the entire repertoire in these donors. Overall, we saw that the average number of binned inputs was >100 for each of the donors, except for the 10x dataset whose average was 51.17 (Fig. 3G). Using this, we calculated the estimated FPR using the power relationship in Fig 3A. We observed that the median FPR of bulk processed data ranged between 18 and 31%, whereas the 10x dataset was higher at 38% (Fig. 3H). Given the number of inputs and the FPR, we then sought to compare the TRACE outputs from the human donor and 10x repertoire to their expected number of calculated false positives. Comparison of the TRACE outputs to the expected number of false positives revealed that each donor generated significantly more TRACE outputs than expected by false positive estimation ($p < 0.0001$). Analysis of the 10x repertoire yielded a similar result ($p < 0.0001$) (Fig. 3I). Summed analysis of the TRACE outputs from bulk ($n = 24{,}230$) to the total expected number of false positives ($n = 3{,}933$) yielded an overall calculated FPR of 16.2%. Within the 10x group, the total number of TRACE outputs was much lower ($n = 1{,}014$), with a proportionally larger expected number of expected false positives ($n = 557$), yielding an overall calculated FPR of 54.9%. Although striking, this increase in FPR in the 10x group was expected considering the distinctly lower total number of sequences in the 10x set compared with the bulk set ($n = 15{,}268$ and 264,548, respectively). To control for false positives in subsequent analyses we created modeled sets for each donor, including the 10x set, that are reflective of the overall background of the TRACE methodology, are matched to each dataset, and are composed entirely of false positives. Therefore, even within the 10x sample, we can assess for signal despite a considerably high false-positive background.

### *Somatically mutated IGHV sequences receive templated tracts from inter- and intrachromosomal templates*

After confirming that TRACE outputs were significantly above expected false-positive background, we turned to characterizing these TRACE outputs ($n = 25{,}244$). First, we mapped the unique locations identified as donor tracts and observed hits from scattered regions around the genome (Fig. 4A–C). TRACE outputs were primarily found intrachromosomally on chromosome 14 as well as interchromosomally on chromosomes 15 and 16 and were statistically enriched for donor tracts on these chromosomes as determined by permutational analysis (Stouffer's Z score [$Z_s$] = 82.78, $p < 0.0001$; $Z_s = 9.26$, $p < 0.0001$; $Z_s = 16.57$, $p < 0.0001$) (Fig. 4C). Notably, we did not observe any TRACE output on chromosome 9 between all patients.

We next assessed whether TRACE outputs cluster at the nucleotide level. First, we identified the fraction of TRACE outputs that cluster within each sample and compared this to the clustered fraction in the matched modeled set. For each sample we found an overall increase in the number of TRACE outputs that cluster ($p = 0.016$) (Fig. 4D). Next, we calculated the number of TRACE outputs that cluster between samples, with at least one other sample possessing an output within 1 kb. We found that 39.4% of TRACE outputs from the original samples clustered within 1 kb as compared with

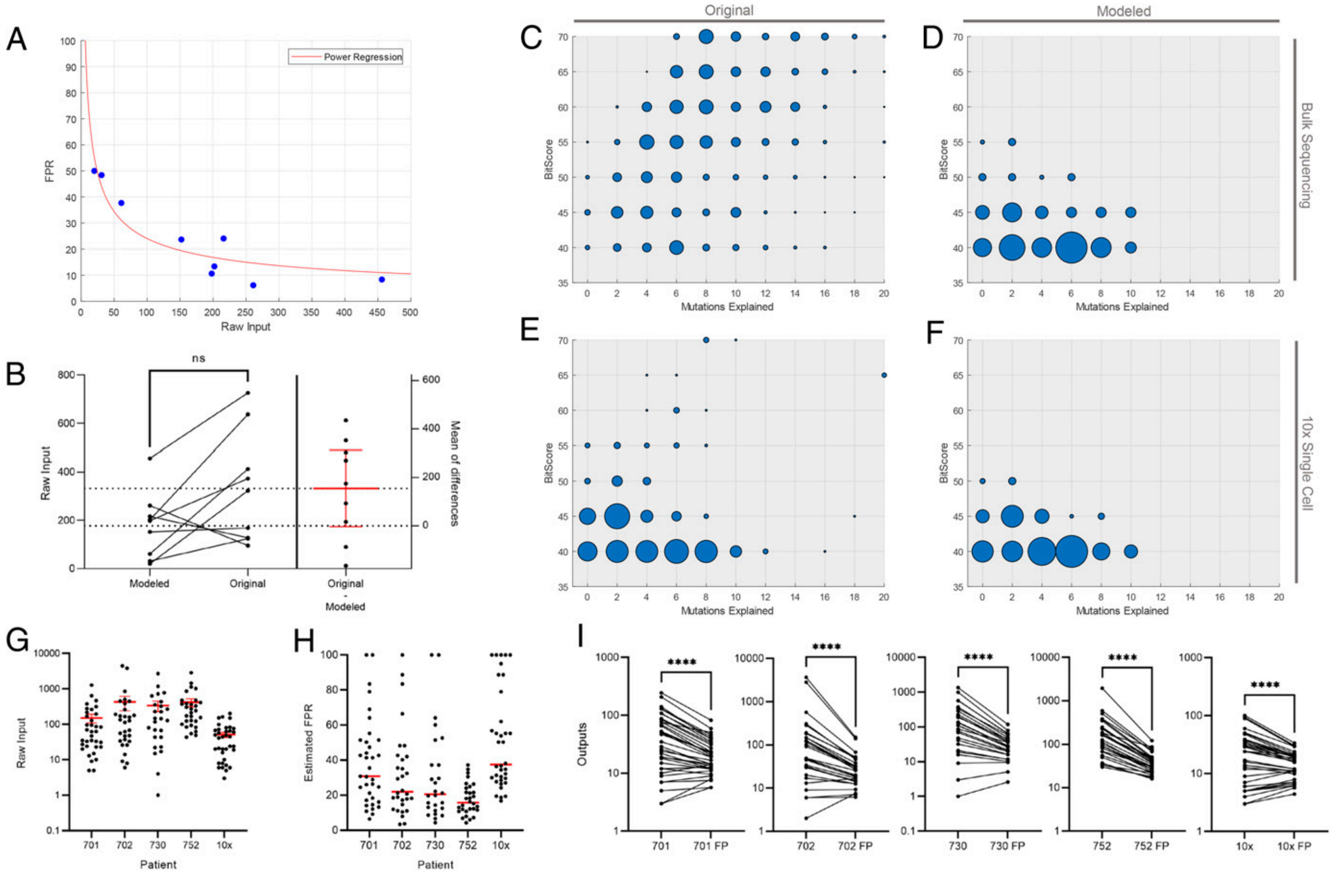
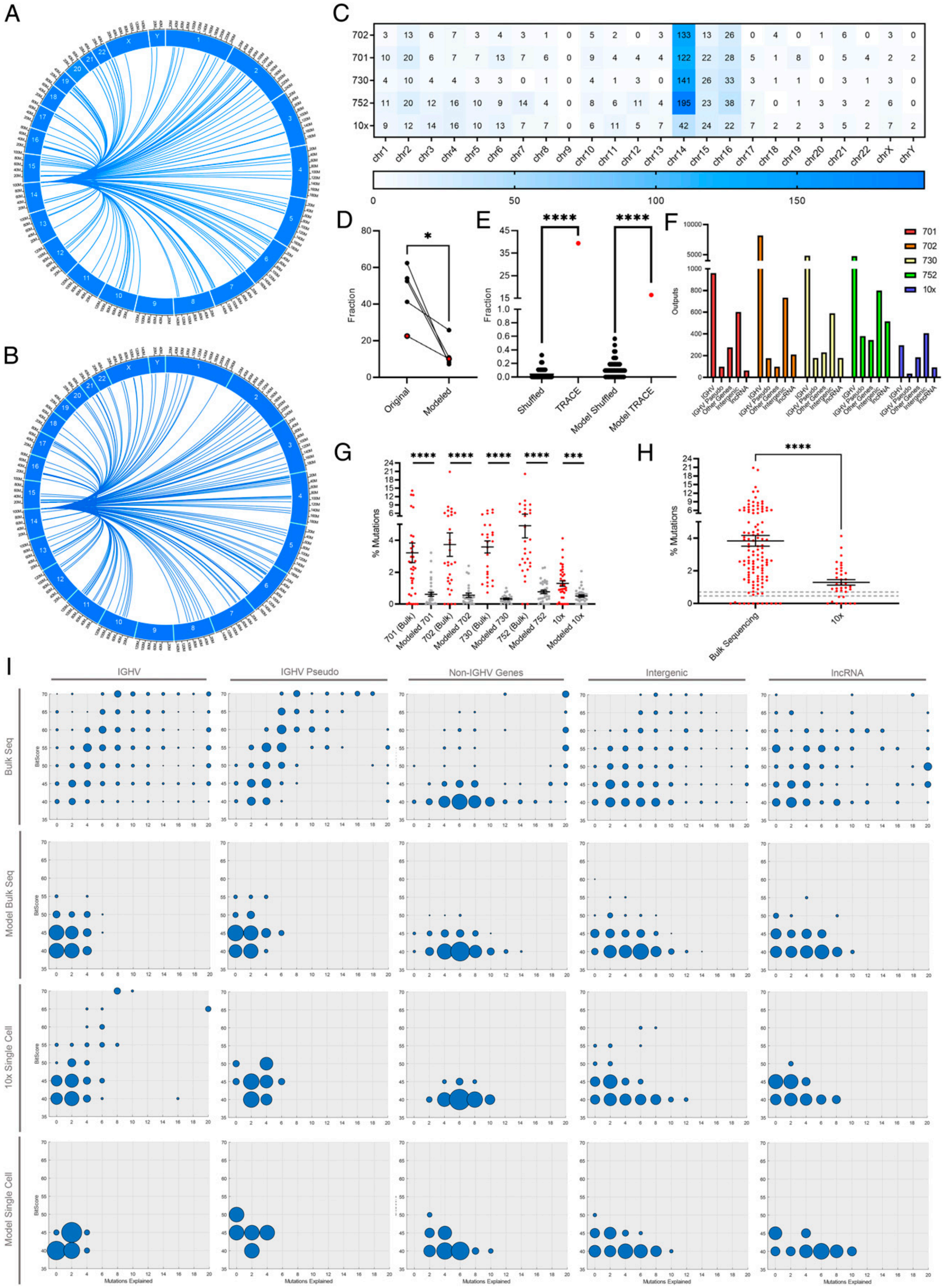


**FIGURE 3.** False-positive analysis of TRACE script indicates minimal contribution of false positives and differences in TRACE output quality. (**A**) Power regression analysis of FPR as a function of TRACE inputs. Dots represent data from three IGHV datasets across three human donors. Modeled data used in the TRACE analysis were treated as a sequence set for TRACE analysis. FPR was calculated as the percent of resultant outputs as compared with inputs before the bin analysis and cleaning step. (**B**) Matched comparison of inputs before bin analysis/cleaning of modeled data used to calculate FPR and corresponding data from human donor samples. Differences between each paired set are shown at right. Mean and SEM are shown in red. (**C**–**F**) Summary data of TRACE outputs from modeled data (i.e., false positives) and from matched original datasets. Depicted are data from bulk sequencing and its corresponding modeled data (C and D) as well as data from single cells and their corresponding modeled data (E and F). Due to the relatively low number of single cell data, (E) and (F) depict summary data for all IGHVs in the sample. (**G**) Shown are inputs for somatically mutated IGHV datasets in four bulk sequenced patients as well as 10x single0cell data. Means for each are shown in red with error bars representing SEM. (**H**) Calculated FPR of each dataset analyzed in (G) from power regression shown in (A). The median for each set is shown in red. (**I**) Comparison of TRACE outputs from all datasets from each of the five datasets to corresponding predicted numbers of false positives expected based on each dataset's estimated FPR. ****$p < 0.0001$; n.s., not significant.

16.5% within the matched modeled sets ($Z_{\text{original}} = 595.18$, $Z_{\text{modeled}} = 175.67$) (Fig. 4E). Although there are significant areas that cluster in the modeled (i.e., false positive) set, a larger fraction (22.9%) of TRACE outputs cluster independently of false-positive background, suggesting that TRACE outputs originate from discrete regions of the genome. Characterization of the types of TRACE-identified donors showed that IGHV genes are the primary contributor of mutations observed across all donors ($p < 0.001$). All other types, including IGHV pseudogenes, were not significant. The distribution of donor types is shown in Fig. 4F. Taken together, these results support our earlier observations in the LAIR1 Ab sequences that specific regions of the genome serve as donor sequences and that homology is a major determinant of donor choice.

We next sought to determine the number of mutations that are accounted for by the TRACE methodology. Within the bulk sequences we observed 3.22% (701), 3.75% (702), 3.59% (730), and 4.91% (752) of mutations attributable to TRACE outputs. These were each significantly greater than each of their respective modeled sets ($p < 0.0001$) (Fig. 4G). In the 10x sample, we observed 1.29% of mutations explained by TRACE outputs, which was also significantly higher than that observed in the modeled set ($p < 0.001$). Given the discrepancy between the bulk-processed and single cell–processed samples, we compared these groups and found that bulk-processed samples had a significantly greater percent of mutations explained than did the 10x sample (3.83% versus 1.29%, $p < 0.0001$) (Fig. 4H). Importantly, we calculated the overall background between the modeled sets of each sample and found the background to be 0.582% (95% confidence interval = 0.459%, 0.705%). Considering this background, we concluded that at least 0.71% of mutations within the 10x sample and as much as 3.25% in the bulk sample were attributable to true positive TRACE outputs.

Quality analysis of the TRACE outputs from the bulk samples, the 10x samples, and their respective modeled sets is shown in Fig. 4I. Within the bulk sample, we observed an increase in the bit score and in the number of mutations explained across all TRACE donor types as compared with the modeled bulk sample. In the 10x sample, we observed a similar pattern only among IGHV and intergenic donors, when compared with the modeled 10x sample. Given that the modeled set is comprised of only false positives, we can observe true positives present in both the bulk and 10x samples as defined by their higher bit scores and mutations explained. We also observe TRACE donors from IGHV pseudogenes,



|  | chr1 | chr2 | chr3 | chr4 | chr5 | chr6 | chr7 | chr8 | chr9 | chr10 | chr11 | chr12 | chr13 | chr14 | chr15 | chr16 | chr17 | chr18 | chr19 | chr20 | chr21 | chr22 | chrX | chrY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 702 | 3 | 13 | 6 | 7 | 3 | 4 | 3 | 1 | 0 | 5 | 2 | 0 | 3 | 133 | 13 | 26 | 0 | 4 | 0 | 1 | 6 | 0 | 3 | 0 |
| 701 | 10 | 20 | 6 | 7 | 7 | 13 | 7 | 6 | 0 | 9 | 4 | 4 | 4 | 122 | 22 | 28 | 5 | 1 | 8 | 0 | 5 | 4 | 2 | 2 |
| 730 | 4 | 10 | 4 | 4 | 3 | 3 | 0 | 1 | 0 | 4 | 1 | 3 | 0 | 141 | 26 | 33 | 3 | 1 | 1 | 0 | 3 | 1 | 3 | 0 |
| 752 | 11 | 20 | 12 | 16 | 10 | 9 | 14 | 4 | 0 | 8 | 6 | 11 | 4 | 195 | 23 | 38 | 7 | 0 | 1 | 3 | 3 | 2 | 6 | 0 |
| 10x | 9 | 12 | 14 | 16 | 10 | 13 | 7 | 7 | 0 | 6 | 11 | 5 | 7 | 42 | 24 | 22 | 7 | 2 | 2 | 3 | 5 | 2 | 7 | 2 |

**FIGURE 4.** Somatically mutated populations of Ag-experienced B cell template somatic mutations from intra- and interchromosomal regions. (**A** and **B**) Circos plots of TRACE-identified templates that contribute to the somatic mutation profile of human donors. Circos plots correspond to (A) donor 701 and (B) the 10x single-cell dataset. (**C**) Heat map depicting the number of unique TRACE outputs per human donor per chromosome. Numbers within cells indicate the number of unique TRACE outputs identified for each chromosome. (**D**) Clustering analysis depicting the fraction *(Figure legend continues)*

intergenic regions, and long noncoding RNA. These donors primarily appear within the bulk sample and are perhaps secondary to the increased sample size present. Although PCR crossover might be expected as a confounder in the case of IGHV pseudogenes and, to a lesser extent, long noncoding RNA, it would not be expected in the case of the intergenic regions.

### *TRACE outputs correlate with multiple facets of GC B cell biology*

To validate the findings of TRACE, we sought to determine whether TRACE-identified regions correlated with GC B cell biology. As the presumptive mechanism for these templated mutations is gene conversion, and gene conversion donor choice is influenced by spatial proximity (20), chromatin accessibility (21), and transcription (22), we sought to elucidate whether these sites identified by TRACE are suggestive of those used for gene conversion.

To do so, we first analyzed published Hi-C chromosome conformation capture data from human GC B cells and naive B cells as reported by Bunting et al. (17), wherein it was demonstrated that BCL6, the key transcription factor of the GC program, mediates widespread chromosomal rearrangement. Although the predominant source of TRACE-identified templates originates on chromosome 14, the redundancy of the IgH locus did not allow us to resolve intrachromosomal interactions. Instead, we analyzed interchromosomal contacts between the IgH locus and the rest of the genome. For each human donor, sites of TRACE-identified donor templates were analyzed as a function of distance from sites of interchromosomal interactions with the IgH locus as identified from the Hi-C data. We observed a consistent and significant pattern in which TRACE outputs from bulk samples were closer to sites of interchromosomal interaction in the GC B cells than those of naive B cells ($p < 0.0001$ for each) (Fig. 5A). Furthermore, these TRACE outputs were all significantly closer than those of the matched modeled sets ($p_{701} = 0.005$, $p_{702} = 0.02$, $p_{730} < 0.0001$, $p_{752} = 0.003$). This pattern was durable from 50 kb to 10 Mb from sites of interchromosomal interaction in all four human donors, indicating that TRACE was predictive of the GC B cell chromosomal conformation but, importantly, not of the naive B cell conformation. Additionally, these results suggest an effect independent from false positives in the modeled set. Unfortunately, we did not observe this pattern in the 10x sample, in which there was no significant difference between the sample and its corresponding modeled set ($p = 0.76$) (Fig. 5B). This was not entirely unexpected given the limited number of TRACE outputs found in this sample coupled with its high FPR. To investigate whether the signal from the 10x sample was similar to that observed in the bulk sequences, we reanalyzed the groups as follows: 10x sample, all bulk samples, and all modeled samples. In doing so, we aimed to identify whether the 10x sample was similar to the combined bulk samples and significantly different from the combined modeled (Fig. 5C). We found that the 10x sample was significantly closer to GC B cell sites of interchromosomal interaction than the merged modeled sample ($p = 0.0006$) and was not significantly different from that the merged bulk sample ($p = 0.23$). This is suggestive, but not indicative, that the 10x sample was also able to identify sites of interchromosomal interaction.

Next, we investigated whether TRACE-identified templates were located within proximity of open chromatin. We assessed genome-wide chromatin accessibility by analyzing assay for transposase-accessible chromatin sequencing data on subsets of B lymphocytes (23) and queried whether TRACE donor templates were located in close proximity (within ≤1 kb) to open chromatin peaks. For each bulk donor, we found that TRACE outputs were significantly enriched within 1 kb of open chromatin peaks as compared with a simulated null distribution an equal number of TRACE output locations (Fig. 5D). This was also true of the 10x and all but one modeled set (Fig. 5E). Given that a large fraction of TRACE outputs are IGHV gene segments, and that multiple rearrangements exist in a population of B cells resulting in an open chromatin signal at multiple IGHV sites, this was the likely confounder in this analysis. Thus, we further analyzed whether interchromosomal donor templates were within 1 kb of open chromatin peaks. We found that interchromosomal TRACE outputs were also significantly enriched within 1 kb of open chromatin peaks for all bulk samples but were no longer significant for any modeled or the 10x sample (Fig. 5F, 5G). As before, this was expected given the low number of TRACE outputs present in the 10x sample. These results suggests that donor templates used for mutagenesis at the IgH locus are associated with regions of the genome that are accessible.

We next queried whether TRACE donor templates are enriched for genes upregulated in the GC B cells by using RNA-seq data from Bunting et al. (17) If true, this would suggest that templates that are used for mutagenesis are transcriptionally active during the mutagenesis program. We found that a total of 5650 genes were upregulated in GC B cells as compared with naive B cells and 2781 genes were downregulated. Between all five human datasets, we found 293 non-IGHV TRACE-identified donor template–containing genes. Of those, 205 genes were present in the naive B and GC B cell RNA-seq data. TRACE-identified donor genes that were not present in the dataset from Bunting et al. included noncoding RNAs and TCR V regions, which is expected, as the studies in Bunting et al. only examined polyadenylated transcripts. Of the 205 genes present, 78 (38%) were present in the upregulated fraction, 28 (13.7%) were contained in the downregulated fraction, and the remaining 99 (48.3%) were in genes whose expression did not change between naive and GC B cells (Supplemental Table I). By permutational analysis, TRACE overlap with the upregulated genes was significantly enriched above background ($Z = 3.78$, $p < 0.001$) as compared with the downregulated genes ($Z = 0.236$, $p = 0.81$). Analysis of RPKM values demonstrated that TRACE outputs were significantly associated with upregulated genes with greater transcription ($Z = 2.37$, $p = 0.018$). In the modeled set, we found a significant association between the upregulated genes and modeled TRACE outputs ($Z = 4.18$, $p < 0.0001$), suggesting that general background could account for our observations. However, analysis of RPKM associated with modeled TRACE outputs did not have any association ($Z = 0.78$, $p = 0.44$), suggesting that the association between transcription activity and TRACE outputs in the original set was independent of background. Upon further analysis, we found that regions of TRACE donor sequences that cluster between human donors overlap with regions that undergo transcription (Supplemental Fig. 3). Taken together, these findings suggests that donor templates preferentially use actively transcribed genes.

of TRACE outputs that cluster within each dataset. Data are matched to the corresponding modeled dataset. (**E**) Clustering analysis depicting the fraction of TRACE outputs that cluster between each dataset. Clustering between modeled datasets is also shown. Data were shuffled as in Fig. 2J ($n = 1000$). (**F**) Bar plot depicting the frequencies of each type of TRACE output identified template per human dataset. (**G**) Shown are the percent mutations explained by TRACE for each IGHV within each human dataset. Percent mutations are also shown for matched modeled sets. (**H**) Percent mutations explained by TRACE grouped by bulk sequencing versus 10x. Dashed lines represent 95% confidence interval (0.459, 0.705) of the model sets depicted in (G). (**I**) Density plots for bulk-sequenced, single-cell, and respective modeled data stratified along type of TRACE output. Density plots are as shown in Fig. 2C. *$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$.

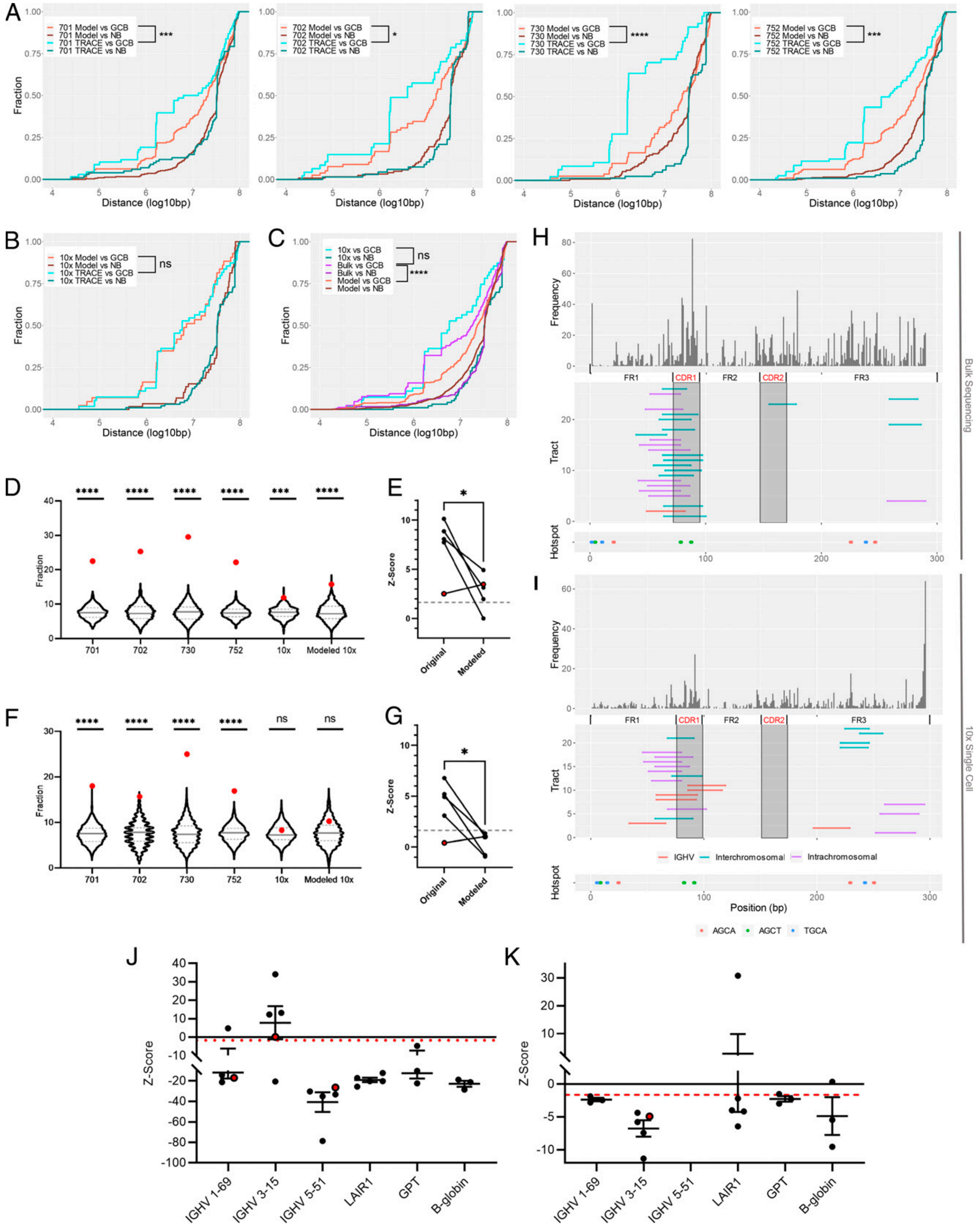


**FIGURE 5.** TRACE hits correlate with multiple aspects of GC B cell biology. (**A**) Cumulative frequency plots of interchromosomal TRACE hits as a function of distance away from interchromosomal Hi-C contact points obtained from Bunting et al. (17). For each bulk human donor (patient) four curves are generated depicting the relationship between either TRACE outputs or modeled TRACE outputs and GC B cell (GCB) Hi-C data and naive B cell (NB) Hi-C data. (**B**) Cumulative frequency plot for TRACE outputs from the 10x single-cell dataset and modeled 10x dataset. (**C**) Cumulative frequency plot for TRACE outputs from summed bulk datasets, the 10x dataset, and summed model datasets. (**D** and **E**) Fraction of TRACE hits that are within 1 kb of open chromatin peaks (red dot). Background data (violin plots) represents 1000 iterations of an equal number of randomly scattered TRACE hits. Heavy dashed line represents the mean, with lighter dashed lines indicating interquartile range. *Z* scores for each dataset are shown in (E). The red data points in (E) represent the 10x data points and the dashed gray line indicates $Z = 1.645$. (**F** and **G**) Fraction of interchromosomal TRACE hits that are within 1 kb of open chromatin peaks. Data are shown as above. *Z* scores are shown in (G). (**H** and **I**) Shown are two plots depicting mutation frequency, location of *(Figure legend continues)*

Having observed multiple biological correlates of TRACE-identified donor sequences, we next sought to analyze the templated mutagenesis recipient sites. We hypothesized that recipient sites would be proximal to sites of palindromic AID hotspots (WGCW), as AID activity is a requirement for gene conversion in B cells (3, 6). We investigated whether such sequences exist within IGHV gene segments and mapped sites that were identified by TRACE as recipients of templated mutagenesis. We found that multiple WGCW sites exist in IGHV gene segments and that, in general, TRACE events map preferentially to regions proximal to these WGCW sites (IGHV1–69: $Z_s = -23.87$, $p < 0.0001$; IGHV3–15: $Z_s = 17.45$, $p = 1$; IGHV5–51: $Z_s = -91.16$, $p < 0.0001$), although this was not unanimous, as in IGHV3–15 (Fig. 5J, Supplemental Fig. 4). Next, we examined nonoverlapping canonical AID hotspots (WRC). We tested whether WRC AID hotspots were associated with TRACE sites. We selected all WRC hotspots on both strands that were not also WGCW sites and found an association between WRC sites and TRACE sites (IGHV1–69: $Z_s = -4.10$, $p < 0.0001$; IGHV3–15: $Z_s = -15.12$, $p < 0.0001$). Unfortunately, not every sequence set could be analyzed in this way, as some had too many WRC sites to simulate a null distribution. Despite this, the association between WRC hotspots and TRACE sites was most pronounced in IGHV3–15, suggesting that in some rearrangements, gene conversion may have preference for WRC rather than WRCW hotspots.

Strikingly, we also observed that TRACE events are primarily clustered in CDR1 and FWR3 regions, across multiple IGHV rearrangements isolated from different human donors, suggesting that diversification of these regions is associated with templated mutagenesis. To rule out any intrinsic effects of the IGHV sequence and GC selection biases, we analyzed the LAIR1 insert (Supplemental Fig. 5), which is a non-IGHV sequence at the IgH locus, as well as the murine passenger transgenes from the Alt laboratory (24) that are both free from selective pressure in addition to bearing no overt homology to the IGHV genes (Supplemental Fig. 6). In these sequences, as well, we found a significant association between TRACE events and the presence of the WGCW (LAIR1: $Z_s = -33.73$, $p < 0.0001$; GPT: $Z_s = -21.58$, $p < 0.0001$; β-globin: $Z_s = -39.59$, $p < 0.0001$) and, to a lesser extent, WRC motifs (LAIR1: $Z_s = 6.24$, $p = 1$; GPT: $Z_s = -3.93$, $p < 0.0001$; β-globin: $Z_s = -8.45$, $p < 0.0001$) (Fig. 5J–K). Taken together, these results suggest a critical role for these motifs for templated mutagenesis and rules out any selection and/or IGHV-specific effects.

## Discussion

In this study, we demonstrate that TRACE is effective at identifying templated mutations and localizing them to distinct regions of the genome. We then validate these predictions by performing a thorough analysis of TRACE's FPR as well as identifying that a subset of TRACE-identified donor sites occurs across individuals, suggesting preferential utilization of certain templates distinct from background. Finally, we show that TRACE analysis of mutation data is correlated with GC B cell biology and is predictive of upregulated genes, chromatin accessibility, and chromosome organization in the nucleus. Taken together, these studies suggest that somatic mutations acquired during the GC reaction are to a very limited degree templated and that such templates can originate from a variety of ectopic sites across the genome.

From the sum of these studies, there is a clear contribution of templated mutagenesis to SHM. This contribution, at best, remains secondary to canonical SHM and only accounts for 0.7% of the mutation load observed. Importantly, however, these events have an affinity for both CDR1 and FR3, suggesting that although the overall contribution of these events is small, they may have important contributions to diversifying the Ag-specific response. Indeed, in the DT40 cell line, which is a model for gene conversion in chickens, CDR1 is the primary site for gene conversion events (6, 25).

Our observation of templated mutation tracts that explain fewer than eight mutations but have a high bit score suggests that templated tracts are also acted upon by canonical SHM, resulting in untemplated mutations. Indeed, observations of gene conversion used in B cell diversification of other species demonstrate that AID activity can take place on existing tracts (8, 26). This further modification of tracts may be an explanation for the discrepancy between the quality of TRACE outputs for LAIR1-containing Abs and those of the human donors and 10x datasets. Given the hypothesized developmental pathway for the LAIR1-containing Ab (19, 27), it is likely that such an Ab is repeatedly and heavily mutated before becoming selected after acquiring the LAIR1 segment. This could lead to "pruning" of some templated mutations within a given tract. Selection of some of these mutations, however, allows TRACE to still identify regions that have undergone this process. This point is particularly supported by detection of templated tracts in LAIR1 in association with hotspot motifs despite the relative decrease in quality of TRACE-identified donors.

By utilizing BLASTn as the mainstay of TRACE, we can readily detect imperfect gene conversion through these obfuscating untemplated events. Given our ability to see these events and their contributed mutations, it would be of great interest to reanalyze the current 5 and 7 k-mer models of SHM (28, 29). In these models, local sequence context is considered to identify preferred sequences for untemplated hypermutation. It is notable that such models, while broadly indicative of susceptibility to mutation, do not fully account for observed hypermutation patterns (30, 31). If templated mutation is occurring, as the evidence in this study suggests, its small contribution to the total mutation load may be a contributing source of residual error in these models.

Importantly, note that the data presented in the present study are generated solely from analysis of mutations and that the predictive capacity of TRACE is derived from nucleotide changes alone. Unlike the other data presented in this study (i.e., assay for transposase-accessible chromatin sequencing, RNA-seq, Hi-C), these types of data are not expected to be representative of higher order biology. That our analyses show that TRACE can predict these higher order

TRACE hits, and location of overlapping AID hotspots for the IGHV5–51 rearrangement from human donor 701 (bulk sequencing) and the 10x single-cell dataset. The mutation frequency plot depicts the frequency of mutation at a given position along the IGHV gene segment, independent of clonality. The plot depicting the location of TRACE hits in recipient sequences depicts each TRACE hit as a horizontal bar indicating its length along the IGHV gene segment. Colors used indicate intrachromosomal (magenta), interchromosomal (blue), and IGHV derived (red). Vertical shaded regions represent CDR1 and CDR2, respectively. Unshaded areas are FR1, FR2, and FR3, respectively. The plot depicting overlapping AID hotspots depicts the location of the WGCW motif. Each color represents a different member of WGCW: AGCT (green), AGCA (red), TGCA (cyan), and TGCT (magenta). (**J** and **K**) *Z* score statistics for Monte Carlo simulations of average shortest distance between TRACE hits and overlapping AID sites (J) or nonoverlapping WRC sites (K) for IGHV1–69, IGHV3–15, and IGHV5–51 as well as LAIR1 insertions from Tan et al. (19) and Pieper et al. (18) and unselected passenger transgenes in a mouse model of GPT and β-globulin reported in Yeap et al. (24). Data not shown in (K) are reflective of *Z* scores that were unable to be calculated. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$.

phenotypes suggests that we are observing a mutational "etching" of DNA reflecting the conditions of the GC B cell stage mediated through the process of templated mutagenesis/gene conversion. This point is further supported by our analysis of the location of TRACE hits within each sequence, demonstrating that there is an association with both the overlapping AID hotspot WRCW and nonoverlapping WRC. In this study, our data suggest that gene conversion is capturing these aspects of GC B cell biology at these sites of AID activity during programmed mutagenesis.

Although these studies do correlate with B cell biology, the exact mechanism responsible for templated mutagenesis is elusive. Homology and proximity appear to be critical mediators, although we have found locations throughout the genome that can serve as templates. Studies by Pieper et al. (18) suggest that B cells are largely unrestricted in accessing templates at distant locations despite preference being given to more proximal regions. Further work on gene conversion in the context of the GC is likely needed to address such questions.

The H and L chain Ig locus in chickens is marked by a single functional full-length V and J segment (32, 33). Following V(D)J recombination, secondary diversification occurs in the bursa of Fabricius in which these single rearrangements are subjected to rounds of gene conversion from an array of Ig pseudogenes located upstream from the rearrangement (33). This secondary round of diversification creates the naive B cell repertoire in the chicken. Gene conversion can also serve to diversify during Ag-specific responses and is known to occur in chicken GC B cells. Similar uses of gene conversion are found in other species known to use the process, including mammals (34–36). In the present work, we demonstrate gene conversion in the circulating repertoire of human donors and, to a very limited extent, in mice. As in chickens, we demonstrate that Ig sequences can be donors for gene conversion events. Importantly, we also demonstrate that non-Ig sequences can serve to diversify B cells, and that these sequences are reflective of conditions within the GC, suggesting that these exogenous sequences are used during the GC reaction. Similar events have been reported within the context of templated insertions (18, 19, 37, 38), but to date, no such events have been implicated as templates for mutagenesis.

Finally, we address the intuitive notion that mutating a sequence will result in a significant match in the genome, owing to the genome's impressive size and complexity. By conducting a thorough series of Monte Carlo experiments in tandem with creating a false-positive set (i.e., the modeled sets), we show that such donors and alignments are unlikely to arise by chance alone. Therefore, we concluded these alignments have biological significance, a point we then test in the latter studies presented in this article. Taken together, these studies provide evidence for a distinct concerted mechanism used to generate additional diversity in human and murine B cells, albeit to a limited degree.

Putting the current study in the context of our earlier studies on templated mutagenesis, we acknowledge the critique that the prior approach by our group was susceptible to a high false-positive rate (12, 13). Therefore, in the current study we have explicitly examined and attempted to minimize the contribution of false positives. As such, the findings in this work should be taken as our group's best estimate of gene conversion to date. However, this work should not be interpreted as a final determination of the rate of gene conversion. Indeed, the major limitation of the TRACE approach presented in the current study is the reliance on highly mutated clusters of mutations, a requirement necessary to determine donor templates at the genomic scale. Our results suggest that many of the templated events at the IgH locus arise from the locus itself. Therefore, future work should be tailored to identifying the local contribution of the locus to mutation patterns. It is likely that TRACE settings could be altered to examine this relatively limited genomic region while still maintaining a low false-positive rate.

Although many questions remain, this work highlights a rather surprising contribution of the larger genome to the generation of diversity in B cells and is complementary to findings from other groups demonstrating the use of exogenous sequences for B cell diversity (18, 19, 37, 38). This process likely has implications on Ab maturation, and further work should be directed at understanding both the mechanism and the functional contribution of templated mutations during an affinity-matured response.

## Acknowledgments

Figures were created with BioRender.

## Disclosures

The authors have no financial conflicts of interest.

## References

1. Victora, G. D., and P. C. Wilson. 2015. Germinal center selection and the antibody response to influenza. *Cell* 163: 545–548.
2. Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102: 553–563.
3. Harris, R. S., J. E. Sale, S. K. Petersen-Mahrt, and M. S. Neuberger. 2002. AID is essential for immunoglobulin V gene conversion in a cultured B cell line. *Curr. Biol.* 12: 435–438.
4. Teng, G., and F. N. Papavasiliou. 2007. Immunoglobulin somatic hypermutation. *Annu. Rev. Genet.* 41: 107–120.
5. Di Noia, J. M., and M. S. Neuberger. 2007. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* 76: 1–22.
6. Arakawa, H., J. Hauschild, and J.-M. Buerstedde. 2002. Requirement of the activation-induced deaminase (AID) gene for immunoglobulin gene conversion. *Science* 295: 1301–1306.
7. Duvvuri, B., and G. E. Wu. 2012. Gene conversion-like events in the diversification of human rearranged IGHV3-23*01 gene sequences. *Front. Immunol.* 3: 158.
8. Tsai, H.-F., N. D'Avirro, and E. Selsing. 2002. Gene conversion-like sequence transfers in a mouse antibody transgene: antigen selection allows sensitive detection of V region interactions based on homology. *Int. Immunol.* 14: 55–64.
9. D'Avirro, N., D. Truong, B. Xu, and E. Selsing. 2005. Sequence transfers between variable regions in a mouse antibody transgene can occur by gene conversion. *J. Immunol.* 175: 8133–8137.
10. Sawyer, S. 1989. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* 6: 526–538.
11. Winstead, C. R., S. K. Zhai, P. Sethupathi, and K. L. Knight. 1999. Antigen-induced somatic diversification of rabbit IgH genes: gene conversion and point mutation. *J Immunol.* 162: 6602–6612.
12. Dale, G. A., D. J. Wilkins, C. D. Bohannon, D. Dilernia, E. Hunter, T. Bedford, R. Antia, I. Sanz, and J. Jacob. 2019. Clustered mutations at the murine and human IgH locus exhibit significant linkage consistent with templated mutagenesis. *J. Immunol.* 203: 1252–1264.
13. Fukuyama, J., B. J. Olson, and F. A. Matsen IV. 2020. Lack of evidence for a substantial rate of templated mutagenesis in B cell diversification. *J. Immunol.* 205: 936–944.
14. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.
15. Brochet, X., M. P. Lefranc, and V. Giudicelli. 2008. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36(Web Server): W503–W508.
16. Tipton, C. M., C. F. Fucile, J. Darce, A. Chida, T. Ichikawa, I. Gregoretti, S. Schieferl, J. Hom, S. Jenks, R. J. Feldman, et al. 2015. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* 16: 755–765.
17. Bunting, K. L., T. D. Soong, R. Singh, Y. Jiang, W. Béguelin, D. W. Poloway, B. L. Swed, K. Hatzi, W. Reisacher, M. Teater, et al. 2016. Multi-tiered reorganization of the genome during B cell affinity maturation anchored by a germinal center-specific locus control region. *Immunity* 45: 497–512.
18. Pieper, K., J. Tan, L. Piccoli, M. Foglierini, S. Barbieri, Y. Chen, C. Silacci-Fregni, T. Wolf, D. Jarrossay, M. Anderle, et al. 2017. Public antibodies to malaria antigens generated by two *LAIR1* insertion modalities. *Nature* 548: 597–601.
19. Tan, J., K. Pieper, L. Piccoli, A. Abdi, M. F. Perez, R. Geiger, C. M. Tully, D. Jarrossay, F. Maina Ndungu, J. Wambua, et al. 2016. A *LAIR1* insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* 529: 105–109.
20. Wang, R. W., C. S. Lee, and J. E. Haber. 2017. Position effects influencing intrachromosomal repair of a double-strand break in budding yeast. *PLoS One* 12: e0180994.
21. Cummings, W. J., M. Yabuki, E. C. Ordinario, D. W. Bednarski, S. Quay, and N. Maizels. 2007. Chromatin structure regulates gene conversion. *PLoS Biol.* 5: e246.

22. Schildkraut, E., C. A. Miller, and J. A. Nickoloff. 2006. Transcription of a donor enhances its use during double-strand break-induced gene conversion in human cells. *Mol. Cell. Biol.* 26: 3098–3105.
23. Scharer, C. D., E. L. Blalock, T. Mi, B. G. Barwick, S. A. Jenks, T. Deguchi, K. S. Cashman, B. E. Neary, D. G. Patterson, S. L. Hicks, et al. 2019. Epigenetic programming underpins B cell dysfunction in human SLE. *Nat. Immunol.* 20: 1071–1082.
24. Yeap, L. S., J. K. Hwang, Z. Du, R. M. Meyers, F. L. Meng, A. Jakubauskaitė, M. Liu, V. Mani, D. Neuberg, T. B. Kepler, et al. 2015. Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* 163: 1124–1137.
25. Nakahara, M., E. Sonoda, K. Nojima, J. E. Sale, K. Takenaka, K. Kikuchi, Y. Taniguchi, K. Nakamura, Y. Sumitomo, R. T. Bree, et al. 2009. Genetic evidence for single-strand lesions initiating Nbs1-dependent homologous recombination in diversification of Ig v in chicken B lymphocytes. *PLoS Genet.* 5: e1000356.
26. Reynaud, C.-A., A. Dahan, V. Anquez, and J.-C. Weill. 1989. Somatic hyperconversion diversifies the single Vh gene of the chicken with a high incidence in the D region. *Cell* 59: 171–183.
27. Robbiani, D. F., and M. C. Nussenzweig. 2016. A new way to diversify antibodies by DNA transposition. *Cell* 164: 601–602.
28. Yaari, G., J. A. Vander Heiden, M. Uduman, D. Gadala-Maria, N. Gupta, J. N. H. Stern, K. C. O’Connor, D. A. Hafler, U. Laserson, F. Vigneault, and S. H. Kleinstein. 2013. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* 4: 358.
29. Elhanati, Y., Z. Sethna, Q. Marcou, C. G. Callan, Jr., T. Mora, and A. M. Walczak. 2015. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370: 20140243.
30. Schramm, C. A., and D. C. Douek. 2018. Beyond hot spots: biases in antibody somatic hypermutation and implications for vaccine design. *Front. Immunol.* 9: 1876.
31. Zhou, J. Q., and S. H. Kleinstein. 2020. Position-dependent differential targeting of somatic hypermutation. *J. Immunol.* 205: 3468–3479.
32. Reynaud, C.-A., V. Anquez, H. Grimal, and J.-C. Weill. 1987. A hyperconversion mechanism generates the chicken light chain preimmune repertoire. *Cell* 48: 379–388.
33. Tang, E. S., and A. Martin. 2007. Immunoglobulin gene conversion: synthesizing antibody diversification and DNA repair. *DNA Repair (Amst.)* 6: 1557–1571.
34. Butler, J. E. 1998. Immunoglobulin diversity, B-cell and antibody repertoire development in large farm animals. *Rev. Sci. Tech.* 17: 43–70.
35. Guo, Y., Y. Bao, Q. Meng, X. Hu, Q. Meng, L. Ren, N. Li, and Y. Zhao. 2012. Immunoglobulin genomics in the guinea pig (*Cavia porcellus*). *PLoS One* 7: e39298.
36. Qin, T., H. Zhao, H. Zhu, D. Wang, W. Du, and H. Hao. 2015. Immunoglobulin genomics in the prairie vole (*Microtus ochrogaster*). *Immunol. Lett.* 166: 79–86.
37. Chen, Y., K. Xu, L. Piccoli, M. Foglierini, J. Tan, W. Jin, J. Gorman, Y. Tsybovsky, B. Zhang, B. Traore, et al. 2021. Structural basis of malaria RIFIN binding by LILRB1-containing antibodies. *Nature* 592: 639–643.
38. Koning, M. T., E. M. Vletter, R. Rademaker, R. D. Vergroesen, I. J. M. Trollmann, P. Parren, C. A. M. van Bergen, H. U. Scherer, S. M. Kiełbasa, R. E. M. Toes, and H. Veelken. 2020. Templated insertions at VD and DJ junctions create unique B-cell receptors in the healthy B-cell repertoire. *Eur. J. Immunol.* 50: 2099–2101.